(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 866 442 A3

# (12) EUROPEAN PATENT APPLICATION

(88) Date of publication A3:
**21.04.1999 Bulletin 1999/16**

(51) Int. Cl.[6]: **G10L 5/06**

(43) Date of publication A2:
**23.09.1998 Bulletin 1998/39**

(21) Application number: **98104599.0**

(22) Date of filing: **13.03.1998**

(84) Designated Contracting States:
**AT BE CH DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE**
Designated Extension States:
**AL LT LV MK RO SI**

(30) Priority: **20.03.1997 US 821349**

(71) Applicant: **AT&T Corp.**
**New York, NY 10013-2412 (US)**

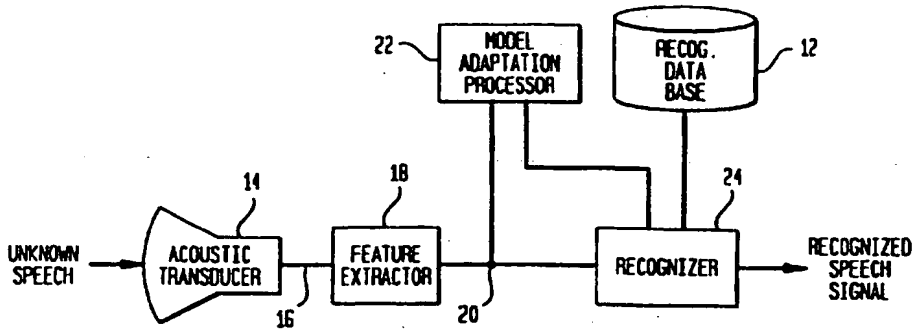
(72) Inventors:
- **Potamianos, Alexandros**
  **Scotch Plains, New Jersey 07076 (US)**
- **Rose, Richard Cameron**
  **Watchung, New Jersey 07060 (US)**

(74) Representative:
**Modiano, Guido, Dr.-Ing. et al**
**Modiano, Josif, Pisanty & Staub,**
**Baaderstrasse 3**
**80469 München (DE)**

## (54) Combining frequency warping and spectral shaping in HMM based speech recognition

(57) Frequency warping approaches to speaker normalization have been proposed and evaluated on various speech recognition tasks. In all cases, frequency warping was found to significantly improve recognition performance by reducing the mismatch between test utterances presented to the recognizer and the speaker independent HMM model. This invention relates to a procedure which compensates utterances by simultaneously scaling the frequency axis and reshaping the spectral energy contour. This procedure is shown to reduce the error rate in a telephone based connected digit recognition task by 30%.

*FIG. 1*




EP 0 866 442 A3

FIG. 5

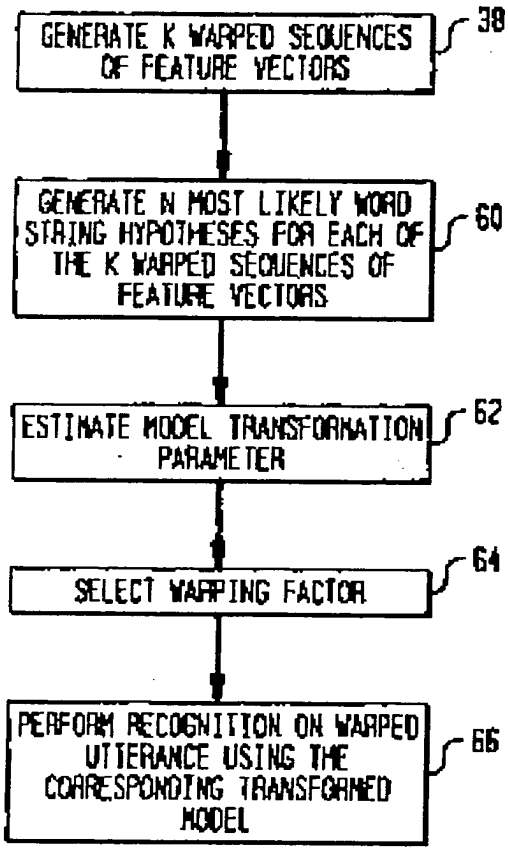


FIG. 3

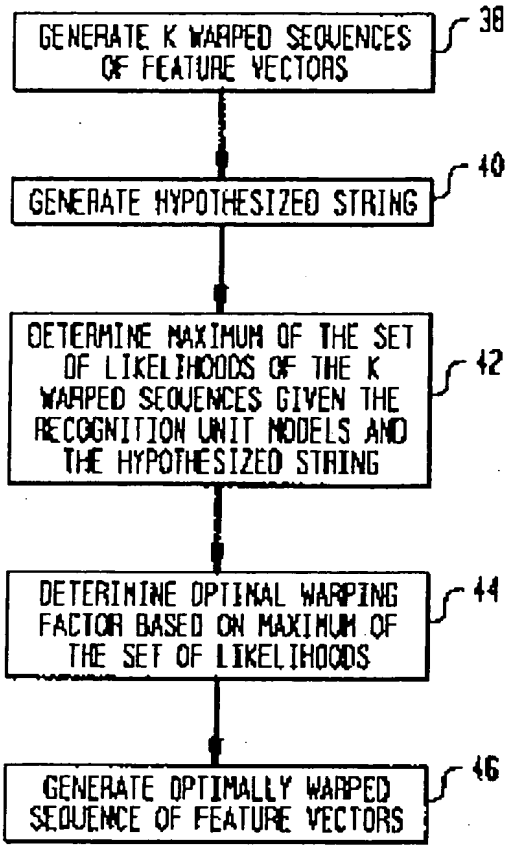
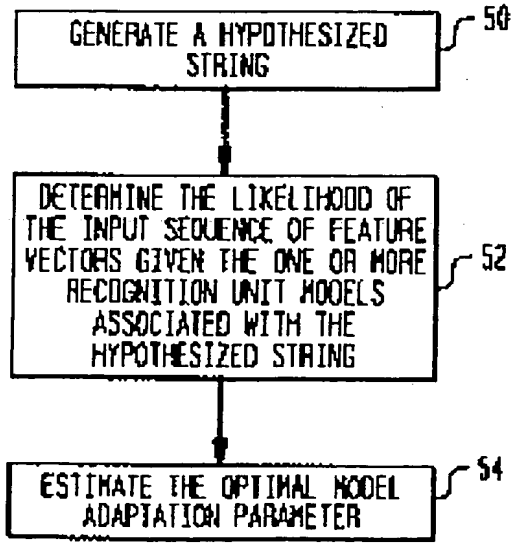



FIG. 4



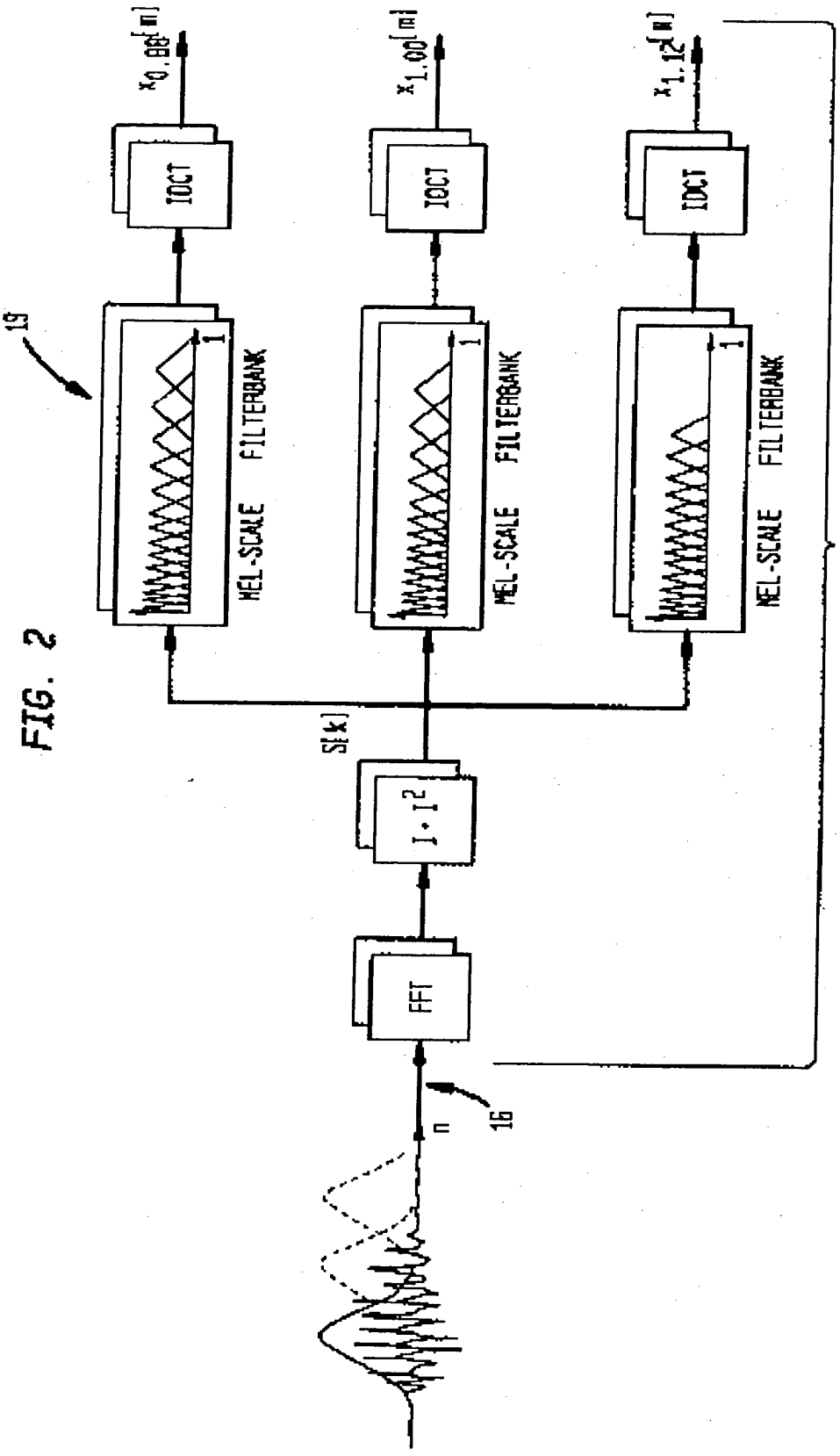

$x_{0.88}[m]$
$x_{1.00}[m]$
$x_{1.12}[m]$
IDCT
IDCT
IDCT
19
1
1
1
MEL-SCALE FILTERBANK
MEL-SCALE FILTERBANK
MEL-SCALE FILTERBANK
18
S(k)
$I \cdot I^2$
FFT
n
16

FIG. 2

FIG. 1

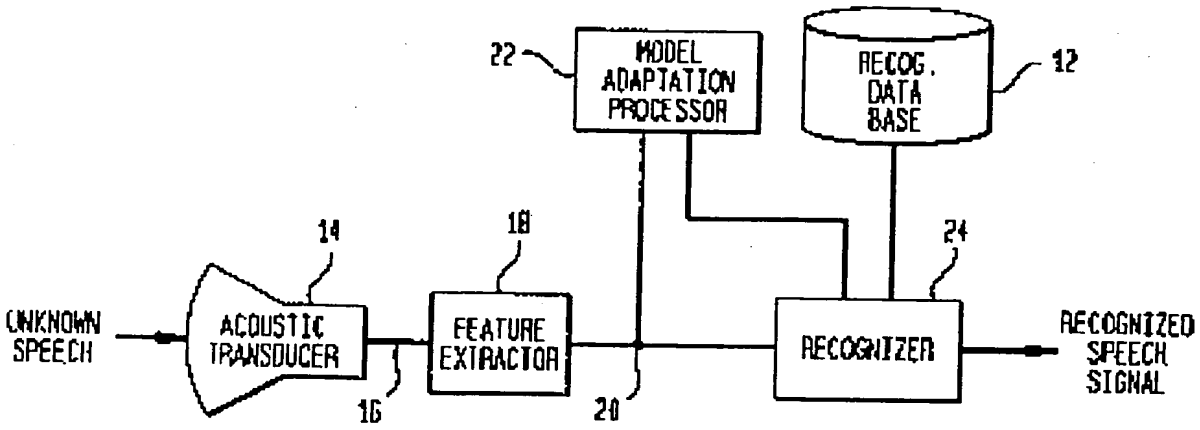


FIG. 5

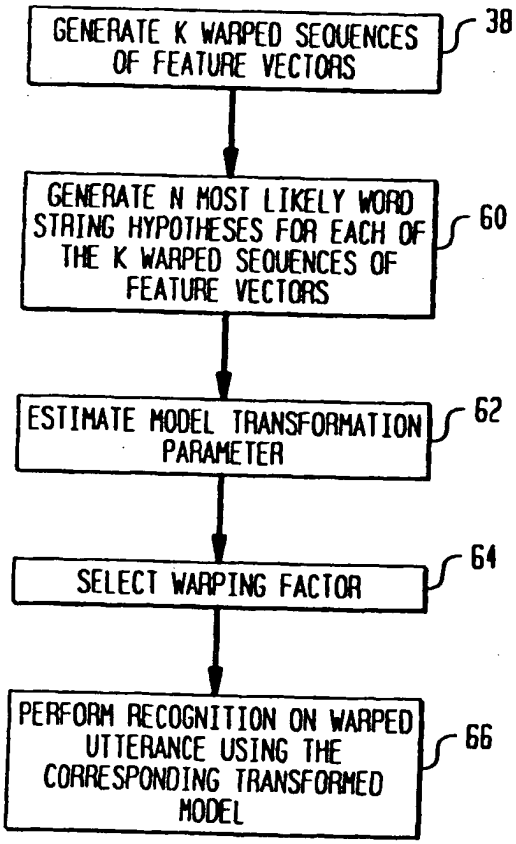


FIG. 3

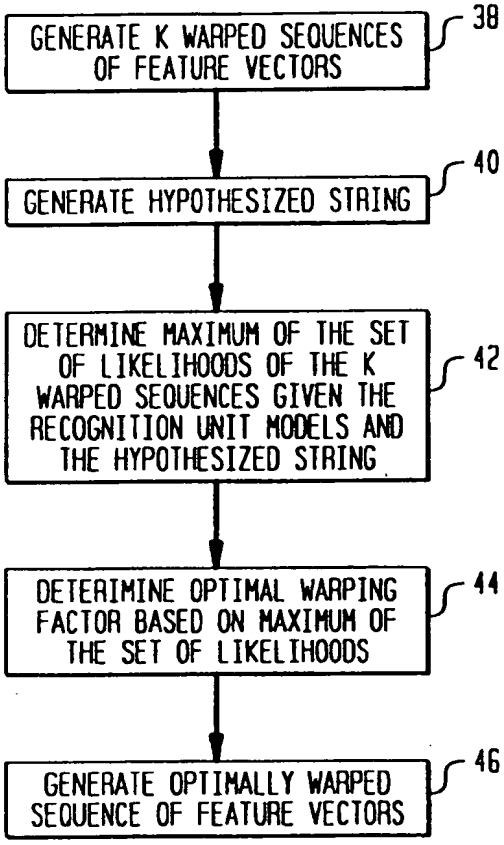
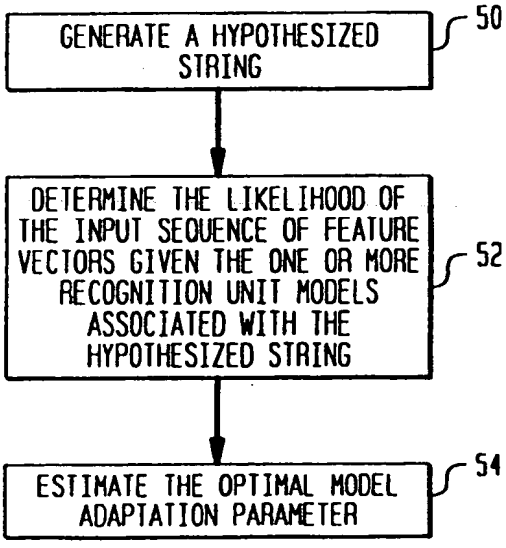



FIG. 4



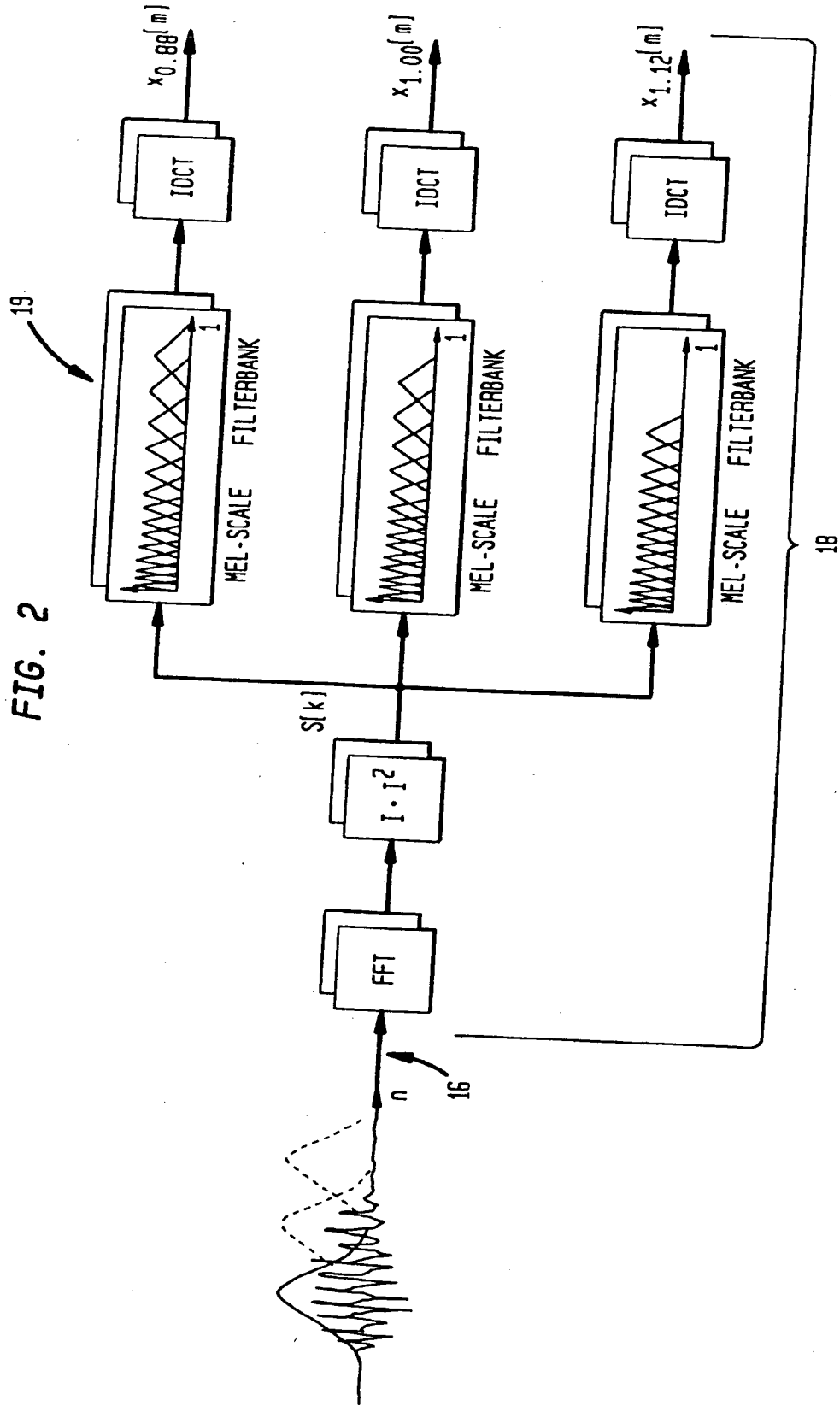
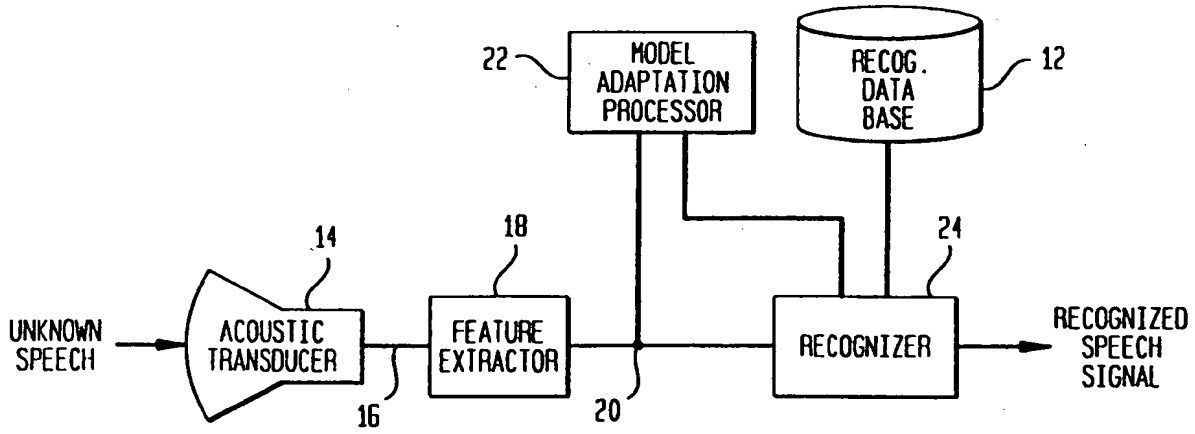



FIG. 2

*FIG. 1*



14. A system as defined in claim 10, wherein:

each recognition unit model including one or more Gaussian distributions,
each of the one or more Gaussian distributions having a mean and a variance, and further comprising a matrix of recognition unit model parameters comprising the mean and the variance of each Gaussian distribution.

15. A system as defined in claim 14, wherein:

the model adaptation processor operating to determine a hypothesized string based on a probabalistic alignment of the warped sequence of feature vectors with respect to the set of recognition unit models, and determine a model adaptation parameter based on the warped sequence of feature vectors and the mean and the variance of a Gaussian distribution of a recognition unit model associated with the hypothesized string,
wherein: the model adaptation parameter in particular corresponds to a shift in the mean of each Gaussian distribution in the set of recognition unit models.

determining a hypothesized string based on a probabalistic alignment of the warped sequence of feature vectors with respect to the set of recognition unit models; and

determining a model adaptation parameter based on the warped sequence of feature vectors and the mean and the variance of a Gaussian distribution of a recognition unit model associated with the hypothesized string;

wherein, when dependent on claim 2, the linear transformation being based on the model adaptation parameter.

**9.** A method as defined in claim 7, wherein step (C) further comprises the step of:

shifting the mean of each Gaussian distribution in the set of recognition unit models based on the model adaptation parameter.

**10.** A speech recognition system, comprising:

an acoustic transducer capable of receiving sound waves representing unknown speech and converting the sound waves into an electrical unknown speech signal;

a feature extractor coupled to the acoustic transducer, wherein the feature extractor generating a warped sequence of feature vectors based on the unknown speech signal, the feature vectors of the warped sequence being warped according to a warping factor;

a memory means for storing a set of recognition unit models;

a model adaptation processor coupled to the feature extractor and the memory means, wherein the model adaptation processor adapting the set of recognition unit models to the unknown speech signal based on the warped sequence of feature vectors; and

a recognizer coupled to the feature extractor and the memory means, wherein the recognizer recognizing the unknown speech signal based on the warped sequence of feature vectors and the set of adapted recognition unit models.

**11.** A system as defined in claim 10, wherein:

the set of recognition unit models comprising one or more hidden Markov models.

**12.** A system as defined in claim 10, wherein:

the feature extractor including a bank of mel-scale filterbanks,

each of the bank of mel-scale filterbanks having a particular spacing and bandwidth of the filters within the mel-scale filterbank corresponding to an amount of frequency transformation and being associated with a particular warping factor,

at least one of the bank of mel-scale filterbanks corresponding to no frequency transformation and being associated with a warping factor of one.

**13.** A system as defined in claim 12, wherein:

the feature extractor operating to

determine an unwarped sequence of feature vectors, which characterizes the unknown speech signal, using the at least one mel-scale filterbank corresponding to no frequency transformation and being associated with a warping factor of one,

determine a hypothesized string based on a probabalistic alignment of the unwarped sequence of feature vectors with respect to the set of recognition unit models,

determine one or more warped sequences of feature vectors, each being warped according to a different warping factor, characterizing the unknown speech signal,

determine, for each of the one or more warped sequences of feature vectors, the likelihood of the warped sequence of feature vectors with respect to one or more recognition unit models associated with the hypothesized string to make a set of likelihoods,

determine the maximum of the set of likelihoods,

select the optimal warping factor based on the maximum of the set of likelihoods, and

identify, based on the optimal warping factor, the warped sequence of feature vectors characterizing the unknown speech signal.

(A) receiving an unknown speech signal;
(B) generating a warped sequence of feature vectors characterizing the unknown speech signal;
(C) adapting a set of recognition unit models to the unknown speech signal; and
(D) recognizing the unknown speech signal based on the warped sequence of feature vectors and the set of adapted recognition unit models.

**2.** A signal processing method for recognizing unknown speech signals, comprising the following steps:

(A) receiving an unknown speech signal;
(B) generating a warped sequence of feature vectors characterizing the unknown speech signal;
(C) applying a linear transformation to the warped sequence of feature vectors; and
(D) recognizing the unknown speech signal based on the linearly transformed warped sequence of feature vectors and the set of recognition unit models.

**3.** A method as defined in claims 1 or 2, wherein:

the set of recognition unit models comprising one or more hidden Markov models.

**4.** A method as defined in claims 1 or 2, wherein step (B) comprises the step of:

providing a bank of mel-scale filterbanks, wherein
each of the bank of mel-scale filterbanks having a particular spacing and bandwidth of the filters within the mel-scale filterbank corresponding to an amount of frequency transformation and being associated with a particular warping factor, and
at least one of the bank of mel-scale filterbanks corresponding to no frequency transformation and being associated with a warping factor of one.

**5.** A method as defined in claim 3, when dependent on claim 1, wherein step (B) further comprises the steps of:

determining an unwarped sequence of feature vectors using the at least one mel-scale filterbank corresponding to no frequency transformation and being associated with a warping factor of one;
determining a hypothesized string based on a probabilistic alignment of the unwarped sequence of feature vectors with respect to the set of recognition unit models;
determining one or more warped sequences of feature vectors, the feature vectors of each warped sequence being warped according to a different warping factor;
determining, for each of the one or more warped sequences of feature vectors, the likelihood of the warped sequence of feature vectors with respect to one or more recognition unit models associated with the hypothesized string to make a set of likelihoods;
determining the maximum of the set of likelihoods;
selecting an optimal warping factor based on the maximum of the set of likelihoods; and
identifying, based on the optimal warping factor, the warped sequence of feature vectors characterizing the unknown speech signal.

**6.** A method as defined in claims 1 or 2, further comprising the step of:

providing a memory means for storing the set of recognition unit models and a matrix of recognition unit model parameters, wherein
each recognition unit model including one or more Gaussian distributions,
each of the one or more Gaussian distributions having a mean and a variance, and
the matrix of recognition unit model parameters comprising the mean and the variance of each Gaussian distribution.

**7.** A method as defined in claim 6, when dependent on claim 1, wherein step (C) comprises the step of:

adjusting one or more recognition unit model parameters.

**8.** A method as defined in claim 6, wherein step (C) comprises the steps of:

TABLE 1 (continued)

| ADAPTATION METHOD | DATA ERROR |
|---|---|
| Baseline + Warp Trained | 2.9% |
| Warp | 2.5% |
| Bias | 2.5% |
| Warp + Bias | 2.2% |
| N-Best + Warp + Bias | 2.1% |

Referring to TABLE 1, the "Baseline" digit accuracy is shown in the first row. The second baseline experiment, labeled "Baseline + Warp Trained", refers to the improved acoustic models obtained by applying frequency warping during training. In such training procedure, first the optimum linear frequency warping factor was estimated for each speaker in the training set so that $P(X^{\alpha}|\alpha,\lambda,H_c)$ was maximized, where $H_c$ is the known transcription (of the speaker's utterances) corresponding to $X$. Then, improved state alignment was obtained using the warped feature vectors $X^{\alpha}$. Finally, HMM models were trained from the original (unwarped) utterances $X$ using the segmentation information obtained from the warped utterances. A 15% reduction in word recognition error rate for the test set was achieved by using warping during training. The "Warp Trained" HMMs are used for the remainder of the adaptation experiments reported in TABLE 1.

In the speech recognition system taught herein, the recognition unit models preferably comprise one or more hidden Markov models (HMMs). The HMMs are trained before testing (i.e., before actual speech recognition in the field). Any conventional training technique to make the set of HMMs can be used in accordance with the principles of the invention. Such training can be iterative and/or discriminative. Recognition unit model training is described in detail in U.S. Patent No. 5,579,436 issued November 26, 1996 to Chou et al., entitled "RECOGNITION UNIT MODEL TRAINING BASED ON COMPETING WORD AND WORD STRING MODELS", previously incorporated by reference herein.

Next, TABLE 1 shows the performance of the speaker adaptation techniques outlined previously when a single utterance is used to estimate the transformation parameters. In the third row of TABLE 1, "Warp", refers to frequency warping such as described previously, in which the amount of linear frequency scaling ranges from 12% compression to 12% expansion and a total of 13 warping factors are used within this range. The fourth row of TABLE 1, "Bias", displays the recognition rate when a single linear bias is estimated for the whole utterance without the use of frequency warping. The optimal bias vector maximizes $P(h_{\gamma}(X)|\gamma,\lambda,H)$, where $H$ is the corresponding transcription (i.e., hypothesis string) obtained from a preliminary decoding pass.

The fifth row of Table 1, labeled "Warp + Bias", refers to frequency warping and linear bias estimation applied in cascade according to the principles of the invention. A separate bias vector $_{\alpha}$ was computed and subtracted from each warped observation sequence $X^{\alpha}$ before the optimal warping index _ was selected as taught herein. Joint optimization of the bias vector and the warping index provides additional performance improvement of the speech recognition system compared to separately optimizing the bias vector and the warping index. The combined optimization of both model transformation and frequency warping is thus shown to obtain a better match between unknown speech utterances and the recognition unit models.

The last row in TABLE 1 labeled "N-Best + Warp + Bias", shows the performance of the combined procedure illustrated by FIG. 5 using the N-best string model generator, in which warping and bias estimation were applied to the top four scoring transcriptions (i.e., the four best hypothesis strings). This form of the combined procedure provides a larger ensemble of models as "starting points" for adaptation, whereby the word recognition error rate can be reduced by approximately 30%. Most of the improvement is due to the combination of the frequency warping and bias adaptation techniques, while some improvement is due to using N-best alternate hypotheses in the process of estimating the transformation parameters. The reduction in error rate obtained by combining frequency warping and spectral shaping in a single process is approximately equal to the sum of the reduction in speech recognition error rates when applying each of the adaptation procedures separately.

While several particular forms of the invention have been illustrated and described, it will also be apparent that various modifications can be made without departing from the spirit and scope of the invention. Where technical features mentioned in any claim are followed by reference signs, those reference signs have been included for the sole purpose of increasing the intelligibility of the claims and accordingly, such reference signs do not have any limiting effect on the scope of each element identified by way of example by such reference signs.

## Claims

1. A signal processing method for recognizing unknown speech signals, comprising the following steps

mal model adaptation parameter, N-best string candidates are used as an additional ensemble of alternatives that can be searched for finding the optimum parameters $\{\alpha,\gamma\}$. The procedure for estimating the optimal parameters $\{\alpha,\gamma\}$ and performing speech recognition is described as follows and with reference to FIG. 5.

(I) Ensemble of Warped Utterances: Given input speech and an ensemble of linear warping factors $\alpha_1, ..., \alpha_K$, generate in step 58 a sequence of warped feature vectors corresponding to each warping factor:

$$X^{\alpha 1}, ..., X^{\alpha K}. \quad (6)$$

(II) N-best Recognition Hypotheses: Generate in step 60 the N most likely word string hypotheses by performing an initial decoding pass for each of the frequency warped utterances:

$$H^1{}_i, ..., H^N{}_i, \text{ for } i = 1, ..., K \quad (7)$$

A conventional N-best string model generator is designed to receive a "best" string model and generate a set of N string models which are highly competitive with the best string model. The N highly competitive string models provide a basis for generating N best strings associated with the N best string models. The N-best string model generator receives hidden Markov model (HMM) parameters from the recognition database and produces a set of string models which are highly competitive with the one or more recognition unit models that best match the input sequence of feature vectors. Determination of the N-best word strings is made through use of DSP implementation of a modified Viterbi decoder. An N-best string model generator that can be used in accordance with the principles of the invention is described in detail in U.S. Patent No. 5,579,436 issued November 26, 1996 to Chou et al., entitled "RECOGNITION UNIT MODEL TRAINING BASED ON COMPETING WORD AND WORD STRING MODELS", which is incorporated by reference as if fully set forth herein.

(III) Estimate Model Transformation: Given HMM model $\lambda$ and N-best string hypotheses $H^n{}_i$, for $n = 1, ..., N$ obtained from each warped sequence of feature vectors $X^{\alpha i}$, compute in step 62 a model transformation vector for each $\alpha_i$:

$$\gamma_i \text{ for } i = 1, ..., K \quad (8)$$

as given by Eq. 5 in order to increase the likelihood

$$P(X^{\alpha i}/\alpha_i,\gamma_i,\lambda,H^n{}_i) \text{ for } i = 1, ..., K \quad (9)$$

(IV) Select Warping Factor: Given $\gamma_i$ obtained for each $\alpha_i$, $i = 1, ..., K$ compute in step 64

$$\{\alpha,\gamma\} = \arg\max_i P(X^{\alpha i}/\alpha_i,\gamma_i,\lambda,H^n{}_i) \quad (10)$$

(V) Recognition: Perform recognition in a final recognition pass on the optimally warped sequence of feature vectors $X^\alpha$ using the corresponding optimally transformed models $h_\gamma(\lambda)$ in step 66.

Single utterance-based adaptation experiments were performed on a connected digit speech corpus that was collected over the public switched telephone network. The training corpus (i.e., known utterances used to make the recognition unit models) consisted of 8802 single to seven digit utterances (a total of 26717 digits), and the test corpus (i.e., unknown utterances to be recognized according to the principles of the invention) contained 4304 utterances (13185 digits) from 242 male and 354 female speakers. Context-independent continuous hidden Markov digit models with mixtures of eight Gaussians per state were used in an exemplary speech recognition system in accordance with the principles of the invention. Speaker adaptation results are displayed in TABLE 1 in terms of the percentage of digits that were erroneously recognized by such exemplary speech recognition system.

TABLE 1

| ADAPTATION METHOD | DATA ERROR |
|---|---|
| Baseline | 3.4% |

where $X^{\alpha}(t)$ is the cepstrum feature vector at time instant $t$ warped by the warping function $g_{\alpha}()$.

To estimate the optimal model adaptation parameter, which in this specific example is a single linear bias $\gamma$ applied to the warped sequence of feature vectors, it was assumed that only the highest scoring Gaussian in the mixture contributes to the likelihood computation. The estimate is thus simplified as follows:

$$\gamma = (\Sigma\ (X^{\alpha}(t) - \mu_j(t))/\sigma_j(t))/(\Sigma\ 1/\sigma_j(t)) \quad (5)$$

where $\mu_j(t)$ and $\sigma_j(t)$ are the mean and variance of the most active Gaussian $j$ in the mixture at time instant $t$.

An exemplary frequency warping technique which can be implemented in the feature extractor is illustrated by the flow diagram of FIG. 3. K warped sequences of feature vectors are generated in step 38, wherein the K warped sequences are each warped by a warping factor, $\alpha$, which is in a range from .88 to 1.12, where a warping factor of 1.00 corresponds to no warping. Warping corresponds to a frequency transformation of the frequency axis in the mel-scale filterbank. The warping factor is an index to a bank of mel-scale filterbanks, each corresponding to a different amount of frequency transformation.

In an initial recognition pass, the unwarped sequence of feature vectors is scored against the set of recognition unit models. A hypothesized string is generated in step 40 based on the highest score from the probabilistic alignment of the unwarped sequence of feature vectors with the recognition unit models.

The K warped sequences of feature vectors (including one sequence warped with a warping factor of 1.00) are scored by determining the likelihood Of the particular warped sequence given the recognition unit models and the hypothesized string in order to make a get of likelihoods. The maximum of the get of likelihoods is selected in step 42, and the warping factor $\alpha$ associated with the warped sequence of feature vectors that scored the maximum likelihood is selected in step 44 as the optimal warping factor, $\alpha$. The optimally warped sequence of feature vectors, in which the feature vectors are each warped according to the optimal warping factor (which is associated with a particular one of the bank of mel-scale filterbanks), is generated in step 46.

The model adaptation processor 22 (FIG. 1) receives a sequence of feature vectors 20 as input and affects the HMM parameters stored in the recognition database 12. An exemplary model adaptation technique which can be implemented in the model adaptation processor 22 is illustrated by the flow diagram of FIG. 4. The model adaptation processor performs a maximum likelihood estimation of model adaptation parameter $\gamma$, which is a set of transformations, where $\mu_i' = \mu_i - \gamma$ for all mean vectors.

Referring to FIG. 4, in an initial recognition pass the sequence of feature vectors is scored against the recognition unit models to generate in step 50 a hypothesized string, H. The role of the hypothesized string H and the probabilistic alignment is to assign feature vectors $X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(T)}$ to model states $j_{(1)}, j_{(2)}, j_{(3)}, \ldots, j_{(T)}$.

For a set of HMMs having one Gaussian distribution per model state, there will be for each model state an associated mean $\mu_{j(t)}$ and an associated variance $\sigma_{j(t)}$, respectively: $\mu_{j(1)}, \mu_{j(2)}, \mu_{j(3)}, \ldots, \mu_{j(T)}$; and $\sigma_{j(1)}, \sigma_{j(2)}, \sigma_{j(3)}, \ldots, \sigma_{j(T)}$.

The likelihood of the input sequence of feature vectors given the one or more recognition unit models associated with the hypothesized string is determined in step 52. The optimal model adaptation parameter is the one associated with the maximum of the likelihood of the sequence of feature vectors given the one or more recognition unit models associated with the hypothesized string. Assuming that only the highest scoring Gaussian in the mixture contributes to the likelihood computation, the optimal model adaptation parameter is estimated in step 54 according to Eq. 5. The optimal model adaptation parameter defined by Eq. 5 reflects a linear transformation to the means of the model states that are assigned to the input sequence of feature vectors. The linear transformation corresponds to a single linear bias which shifts the means of the model states.

Because of the simplicity of the linear transformation (i.e., a single linear bias) applied to the models, the inverse of such transformation, in the form of a single linear bias, can be applied directly to the input sequence of feature vectors, as discussed previously. Thus, in a specific embodiment of the invention, the optimal model adaptation parameter, $\gamma$, is used to adapt the recognition unit models to the unknown input utterance (characterized by the input sequence of feature vectors) by applying the inverse of the transformation corresponding to $\gamma$ to the sequence of feature vectors. This inverse transformation is somewhat more easily implemented.

In a second recognition pass, speech is recognized by scoring the linearly biased sequence of feature vectors in a probabilistic alignment with the set of HMMs to generate a recognized speech signal.

FIG. 5 presents a flow diagram for describing joint optimization of the parameters of frequency warping and model adaptation in a maximum likelihood framework in a specific embodiment of the invention. The form of the procedure for simultaneous estimation of spectral warping and spectral shaping functions is dictated largely by the fact that it is sufficient to select an optimal warping function from a relatively small ensemble of possible warping functions. As a result of this fact, a sequence of warped feature vectors $X^{\alpha i}$ can be generated for each warping factor $\alpha_i$ for i = 1 to K, and these warped sequences of feature vectors can be used to solve the closed form expression given in Eq. 5 for a corresponding model transformation vector $\gamma_i$. The optimal parameters $\{\alpha,\gamma\}$ are then chosen based on the likelihood of the warped utterance with respect to the most likely transformed model. In obtaining the optimal warping factor and the opti-

DSP32C, read-only memory (ROM) for storing software performing the operations discussed below, and random access memory (RAM) for storing DSP results. Very large scale integration (VLSI) hardware embodiments, as well as custom VLSI circuitry in combination with a general purpose DSP circuit, may also be provided. Use of DSPs is advantageous since the signals processed represent real physical signals and processes, such as speech signals, room background noise, etc.

Combining a parametric linear transformation of the recognition unit models and a parametric frequency warping of the features extracted from the unknown input utterance in the telephone-based speech recognition system according to the principles of the invention expands the ensemble of alternatives evaluated during model adaptation to obtain a better match between the input utterance and the recognition unit models.

Referring to FIG. 1, a speech recognition system according to the principles of the invention includes a recognition database 12 for storing recognition unit models which comprise one or more HMMs and HMM parameters associated therewith. Each of the one or more HMMs includes one or more (e.g., eight) Gaussian distributions per state, each Gaussian distribution having a mean and a variance (which are referred to as model parameters). The speech recognition system shown in FIG. 1 includes an acoustic transducer 14, such as a microphone in the handset of a telephone, for receiving audible sound waves representing unknown speech caused by expansion and rarefication of molecules of air with associated impurities. The acoustic transducer 14 converts the sound waves into electrical unknown speech signals 16. A feature extractor 18 is in electrical connection with the electrical signal output 16 of the acoustic transducer 14. Referring to FIG. 2, the feature extractor 18 comprises, at least in part, a bank of K mel-scale filterbanks 19, K being an integer. Each of the bank of mel-scale filterbanks is associated with a warping factor $\alpha$ in a range from .88 to 1.12 and corresponds to an amount of frequency transformation in the frequency domain. A warping factor of 1.00 means no frequency transformation to the spacing and bandwidth of the filters within the mel-scale filter. Referring to FIG. 1, the feature extractor 18 identifies an optimally warped sequence of feature vectors 20 best characterizing the electrical unknown speech signal 16. A model adaptation processor 22 is coupled to the recognition database 12, the feature extractor 18, and a recognizer 24. The recognizer 24 is coupled to the recognition database 12. In an illustrative embodiment of the invention, the model adaptation processor 22 adapts the parameters of the HMMs (i.e., the means and/or variances of the Gaussian distributions) stored in the recognition database 12 to the unknown speech signal 16 by applying a linear transformation to the means of the Gaussian distributions of the HMMs based on the sequence of feature vectors 20 and output from the recognizer 24. The recognizer 24 compares a plurality of the adapted HMMs with the warped sequence of feature vectors 20 to determine a comparison score for each such model, selects the highest comparison score, and generates a recognized speech signal based on the highest score.

Speech recognition performance is improved according to the principles of the invention by combining frequency warping and model adaptation and jointly optimizing the parameters of frequency warping and model adaptation in a maximum likelihood framework. The optimal parameters of the model transformation $\gamma$ and the frequency warping _ can be simultaneously estimated so that:

$$\{\_,\gamma\} = \arg\max_{\{\alpha,\gamma\}} P(X^{\alpha} \mid \alpha,\gamma,\lambda_{\gamma}, H) \qquad (3)$$

The combined optimization procedure is described by Eq. 3. In view of Eq. 3, $h_{\gamma}()$ is a set of transformations applied to the means of the recognition unit model distributions or, as described subsequently, to the observation sequence in the context of speaker adaptation from single utterances in the exemplary speech recognition system.

A computationally efficient implementation of the combined optimization procedure can be used for simple definitions of the model transformation, $h_{\gamma}()$. If the model transformation corresponds to a single fixed (i.e., data independent) transformation applied to all the HMM means, then it can be applied to the observation sequence 20 (FIG. 1) instead of the HMMs; in that case, however, the inverse of the transformation is applied to the sequence of observations 20. Herein, for simplicity, the same notation is used to describe transformations applied to either the sequence of cepstral observations (i.e., the sequence of feature vectors 20) or to the HMMs stored in the recognition database 12.

Frequency warping is applied by feature extractor 18 directly on the cepstral observation sequence during testing. This simplifies significantly both the computational load and the memory requirements of the combined frequency warping and model adaptation procedure taught herein.

Single utterance-based adaptation experiments were performed on a connected digit speech corpus that was collected over the public switched telephone network. In such adaptation experiments, the combined process included linear frequency warping of the sequence of feature vectors followed by a single linear bias applied to such warped sequence of feature vectors. Such combined process can be described as follows:

$$h_{\gamma}(X^{\alpha}(t)) = X^{\alpha}(t) - \gamma, \qquad (4)$$

the means of the HMMs. In view of that reference, frequency warping for speaker normalization and a linear transformation in the cepstral feature space are the same, and would therefore have equivalent effects. McDonough et al. teaches that combining frequency warping for speaker normalization and a linear transformation in the cepstral feature space are redundant.

## SUMMARY OF THE INVENTION

We have discovered that the ineffectiveness of frequency warping for speaker normalization for a particular subset of utterances is due to the interaction of various sources of variability in speech recognition performance in the process of estimating the "best" warping function. If both model adaptation and frequency warping for speaker normalization are limited by the initial relationship between the HMMs and the input utterance, then the solution to this problem is to search for an optimal warping function and an optimal model transformation in the same procedure.

Since a spoken utterance may be simultaneously affected by many sources of speech recognition performance variability, and since there may be many acoustic correlates associated with a given source of variability, it is important that different procedures for compensating for acoustic distortions be tightly coupled with one another. According to the principles of the invention, linear model transformation and frequency warping are implemented as a single combined procedure in an HMM-based speech recognition system to compensate for these sources of speech recognition performance variability.

In an illustrative embodiment of the invention, unknown speech in the form of sound waves is received by an acoustic transducer and is converted into an electrical unknown speech signal. An optimally warped sequence of feature vectors is determined which characterizes the unknown speech signal, each of the feature vectors in the sequence being warped according to an optimal warping factor. Recognition unit models stored in a memory are adapted to the unknown speech signal. A plurality of the adapted recognition unit models is compared with the optimally warped sequence of feature vectors to determine a comparison score for each such model. The highest comparison score is selected and the unknown speech is recognized based on the highest score.

Combining frequency warping of the cepstrum observation sequence and linear transformation of the recognition unit models improves speech recognition performance substantially more than when using either of these techniques alone, contrary to the teachings of "An Approach To Speaker Adaptation Based On Analytic Functions" by McDonough et al. The combined procedure according to the principles of the invention successfully compensates for mismatches between speaker populations used for training and speaker populations encountered during testing in HMM-based speech recognition.

Other aspects and advantages of the invention will become apparent from the following detailed description and accompanying drawing, illustrating by way of example the features of the invention.

## BRIEF DESCRIPTION OF THE DRAWING

In the drawing:

FIG. 1 is a schematic view of a speech recognition system in accordance with the principles of the invention;
FIG. 2 is a schematic view of a bank of mel-scale filterbanks for use in the speech recognition system depicted in FIG. 1;
FIG. 3 is a flow diagram for describing frequency warping in accordance with the principles of the invention;
FIG. 4 is a flow diagram for describing model adaptation in accordance with the principles of the invention; and
FIG. 5 is a flow diagram for describing a joint optimization process in accordance with the principles of the invention.

## DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

For a better understanding of the invention, together with other and further objects, advantages, and capabilities thereof, reference is made to the following disclosure and the figures of the drawing, where like reference characters designate like or similar elements.

For clarity of explanation, the illustrative embodiments of the present invention are presented as comprising individual functional blocks (including functional blocks labeled as "processors"). The functions these blocks represent may be provided through the use of either shared or dedicated hardware, including, but not limited to, hardware capable of executing software. For example, the functions of processors presented in the figures of the drawing may be provided by a single shared processor. (Use of the term "processor" should not be construed to refer exclusively to hardware capable of executing software.)

Illustrative embodiments may comprise digital signal processor (DSP) hardware, such as the AT&T DSP16 or

quency warping does not improve speech recognition performance.

The frequency warping approach to speaker normalization compensates mainly for inter-speaker vocal tract length variability by linear warping of the frequency axis (i.e., applying a frequency transformation to the frequency axis in the frequency domain), by a factor $\alpha$, where an $\alpha = 1.00$ corresponds to no warping (no frequency transformation).

The front-end of a conventional speech recognizer processes samples of an unknown speech signal. The samples are obtained from recording windows of a specified duration (e.g., 10 ms) on the unknown speech signal, and such windows may overlap. The samples of the unknown speech signal are processed using a fast Fourier transform (FFT) component. The output of the FFT component is further processed and coupled to a mel-scale filterbank (which is also referred to as a mel-cepstrum filterbank). The mel-scale filterbank is a series of overlapping bandpass filters, wherein the bandpass filters in the series have a spacing and bandwidth which increases with frequency along the frequency axis. The output of the mel-scale filterbank is a spectral envelope. An additional transformation to the spectral envelope provides a sequence of feature vectors, X, characterizing the unknown speech signal.

In previous practice, frequency warping has been implemented in the mel-scale filterbank of the front-end of the speech recognizer by linear scaling of the spacing and bandwidth of the filters within the mel-scale filterbank. The warping factor is an index to the amount of linear scaling of the spacing and bandwidth of the filters within the mel-scale filterbank. Scaling the mel-scale filterbank in the front-end is equivalent to resampling the spectral envelope using a compressed or expanded frequency range. Changes in the spectral envelope are directly correlatable to variations in vocal tract length.

In frequency warping for speaker normalization according to previous practice, an ensemble of warping factors is made available, each being an index corresponding to a particular amount of linear scaling, and thus, to a particular spacing and bandwidth of the filters within the mel-scale filterbank. For each utterance, the optimal warping factor _ is selected from a discrete ensemble of possible values so that the likelihood of the warped utterance is maximized with respect to a given HMM and a given transcription (i.e., a hypothesis of what the unknown speech is). The values of the warping factors in the ensemble typically vary over a range corresponding to frequency compression or expansion of approximately ten percent. The size of the ensemble is typically ten to fifteen discrete values.

Let $X^{\alpha} = g_{\alpha}(X)$ denote the sequence of cepstral observation vectors (i.e., the sequence of feature vectors), where each observation vector (i.e., each feature vector) is warped by the function $g_{\alpha}()$, and the warping is assumed to be linear. If $\lambda$ denotes the set of HMMs and the parameters thereof, the optimal warping factor is defined as:

$$\_ = \arg\max_{\alpha} P(X^{\alpha} \mid \alpha, \lambda, H) \qquad (1)$$

where $H$ is the transcription (i.e., a decoded string) obtained from an initial recognition pass using the unwarped sequence of feature vectors X. This frequency warping technique is computationally efficient since maximizing the likelihood in Eq. 1 involves only the forced probabilistic alignment of the warped observation vectors $X^{\alpha}$ to a single string H. Finally, the frequency-warped sequence of feature vectors $X^{\alpha}$ is used in a second recognition pass to obtain the final recognized string.

Frequency warping for speaker normalization according to previous practice transforms an utterance according to a parametric transformation, $g_{\alpha}()$, in order to maximize the likelihood criterion given in Eq. 1.

There is a large class of maximum likelihood-based model adaptation procedures that can be described as parametric transformations of the HMMs. For these procedures, let $\lambda_{\gamma} = h_{\gamma}(\lambda)$ denote the models obtained by a parametric linear transformation $h_{\gamma}()$ of the original set of HMMs and parameters thereof. The form of the parametric linear transformation can depend on the nature of the sources of non-uniformity in speech recognition performance and the number of observations (i.e., the sequence of feature vectors) available for estimating the parameters of the transformation.

A maximum likelihood criterion similar to that used for estimating $\alpha$ is used for estimating $\gamma$:

$$\gamma = \arg\max_{\gamma} P(X \mid \gamma, \lambda_{\gamma}, H) \qquad (2)$$

The article by McDonough et al. entitled "An Approach To Speaker Adaptation Based On Analytic Functions", *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, GA, May 1996, pages 721-724, suggests that in an HMM-based speech recognition system the effect of frequency warping for speaker normalization is equivalent to that obtained from either a linear transformation applied to the cepstral feature space or a linear transformation applied to

## Description

### FIELD OF THE INVENTION

This invention relates to speech recognition systems generally, and more particularly to a signal processing technique which combines frequency warping and spectral shaping for use in hidden Markov model-based speech recognition systems.

### BACKGROUND OF THE INVENTION

Speech recognition is a process by which an unknown speech utterance (usually in the form of a digital PCM signal) is identified. Generally, speech recognition is performed by comparing the features of an unknown utterance to the features of known words or word strings.

The features of known words or word strings are determined with a process known as "training". Through training, one or more samples of known words or strings (training speech) are examined and their features (or characteristics) recorded as reference patterns (or recognition unit models) in a database of a speech recognizer. Typically, each recognition unit model represents a single known word. However, recognition unit models may represent speech of other lengths such as subwords (e.g., phones, which are the acoustic manifestation of linguistically-based phonemes). Recognition unit models may be thought of as building blocks for words and strings of words, such as phrases or sentences.

To recognize an utterance in a process known as "testing", a speech recognizer extracts features from the utterance to characterize it. The features of the unknown utterance are referred to as a test pattern. The recognizer then compares combinations of one or more recognition unit models in the database to the test pattern of the unknown utterance. A scoring technique is used to provide a relative measure of how well each combination of recognition unit models matches the test pattern. The unknown utterance is recognized as the words associated with the combination of one or more recognition unit models which most closely matches the unknown utterance.

Recognizers trained using both first and second order statistics (i.e., spectral means and variances) of known speech samples are known as hidden Markov model (HMM) recognizers. Each recognition unit model in this type of recognizer is an N-state statistical model (an HMM) which reflects these statistics. Each state of an HMM corresponds in some sense to the statistics associated with the temporal events of samples of a known word or subword. An HMM is characterized by a state transition matrix, A (which provides a statistical description of how new states may be reached from old states), and an observation probability matrix, B (which provides a description of which spectral features are likely to be observed in a given state). Scoring a test pattern reflects the probability of the occurrence of the sequence of features of the test pattern given a particular model. Scoring across all models may be provided by efficient dynamic programming techniques, such as Viterbi scoring. The HMM or sequence thereof which indicates the highest probability of the sequence of features in the test pattern occurring identifies the test pattern.

A major hurdle in building successful speech recognition systems is non-uniformity in performance thereof across a variety of conditions. Many successful compensation and normalization techniques have been proposed in an attempt to deal with differing sources of non-uniformity in performance. Some examples of typical sources of non-uniformity in performance in telecommunications applications of speech recognition include inter-speaker, channel, environmental, and transducer variability, and various types of acoustic mismatch.

Model adaptation techniques have been used to improve the match during testing (i.e., during recognition of unknown speech) between a set of unknown utterances and the hidden Markov models (HMMs) in the recognizer database. Some model adaptation techniques involve applying a linear transformation to the HMMs. The parameters of such a linear transformation can be estimated using a maximum likelihood criterion, and then the transformation is applied to the parameters of the HMMs. A perplexing problem not heretofore solved is the existence of speakers in a population for whom speech recognition performance does not improve after model adaptation using such linear transformation techniques. This is especially true for unsupervised, single utterance-based adaptation scenarios.

It is generally thought that only those distributions in the HMMs that are likely to have been generated (during training) by the unknown utterance have a chance to be mapped more closely to the target speaker with such linear transformation model adaptation techniques. Therefore, if the "match" between the HMMs and the unknown utterance is not reasonably "good" to begin with and the number of unknown utterances is limited (such as for example in a single utterance-based adaptation scenario), then the utterance cannot "pull" the model to better match the target speaker in such conventional model adaptation techniques. Thus, there exists a subset of utterances for which model adaptation does not improve speech recognition performance.

Frequency warping for speaker normalization has been applied to telephone-based speech recognition applications. In previous testing practice, frequency warping for speaker normalization has been implemented by estimating a frequency warping function that is applied to the unknown input utterance so that the warped unknown utterance is better matched to the given HMMs. As is the case for model adaptation, there exists a subset of utterances for which fre-

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 866 442 A2

# (12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
**23.09.1998 Bulletin 1998/39**

(51) Int. Cl.[6]: **G10L 5/06**

(21) Application number: **98104599.0**

(22) Date of filing: **13.03.1998**

(84) Designated Contracting States:
**AT BE CH DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE**
Designated Extension States:
**AL LT LV MK RO SI**

(30) Priority: **20.03.1997 US 821349**

(71) Applicant: **AT&T Corp.**
**New York, NY 10013-2412 (US)**

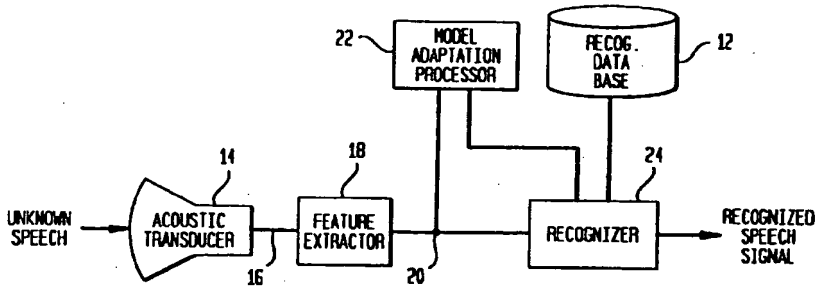
(72) Inventors:
- **Potamianos, Alexandros**
  **Scotch Plains, New Jersey 07076 (US)**
- **Rose, Richard Cameron**
  **Watchung, New Jersey 07060 (US)**

(74) Representative:
**Modiano, Guido, Dr.-Ing. et al**
**Modiano, Josif, Pisanty & Staub,**
**Baaderstrasse 3**
**80469 München (DE)**

## (54) Combining frequency warping and spectral shaping in HMM based speech recognition

(57) Frequency warping approaches to speaker normalization have been proposed and evaluated on various speech recognition tasks. In all cases, frequency warping was found to significantly improve recognition performance by reducing the mismatch between test utterances presented to the recognizer and the speaker independent HMM model. This invention relates to a procedure which compensates utterances by simultaneously scaling the frequency axis and reshaping the spectral energy contour. This procedure is shown to reduce the error rate in a telephone based connected digit recognition task by 30%.

*FIG. 1*




EP 0 866 442 A2

European Patent Office

# EUROPEAN SEARCH REPORT

Application Number
EP 98 10 4599

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.6) |
|---|---|---|---|
| A | ROSE R C ET AL: "A user-configurable system for voice label recognition" PROCEEDINGS ICSLP 96. FOURTH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (CAT. NO.96TH8206), PROCEEDING OF FOURTH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING. ICSLP '96, PHILADELPHIA, PA, USA, 3-6 OCT. 1996, pages 582-585 vol.2, XP002093541 ISBN 0-7803-3555-4, 1996, New York, NY, USA, IEEE, USA * paragraph 5 * --- | 1-4, 10-12 | |
| A | LEGGETTER C J ET AL: "MAXIMUM LIKELIHOOD LINEAR REGRESSION FOR SPEAKER ADAPTATION OF CONTINUOUS DENSITY HIDDEN MARKOV MODELS" COMPUTER SPEECH AND LANGUAGE, vol. 9, no. 2, April 1995, pages 171-185, XP000631203 * paragraph 2 * --- | 1-3,6,7, 9-11,14, 15 | |
| A,D | US 5 579 436 A (CHOU WU ET AL) 26 November 1996 * column 9, line 35 - column 12, line 37 * --- | 1-3,6,7, 9-11,14, 15 | TECHNICAL FIELDS SEARCHED (Int.Cl.6) |
| A | JP 08 211887 A (MITSUBISHI ELECTRIC CORP) 20 August 1996 | 1-3,10, 11 | |
| L | -& US 5 742 928 A (SUZUKI) 21 April 1998 * column 4, line 62 - column 5, line 16 * * column 11, line 51 - column 12, line 43 * * claims 1-3 * ----- | 1-3,10, 11 | |

The present search report has been drawn up for all claims

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| THE HAGUE | 16 February 1999 | Wanzeele, R |

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

European Patent Office

# EUROPEAN SEARCH REPORT

Application Number
EP 98 10 4599

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.6) |
|---|---|---|---|
| X,P | POTAMIANOS A ET AL: "On combining frequency warping and spectral shaping in HMM based speech recognition" 1997 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (CAT. NO.97CB36052), 1997 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, MUNICH, GERMANY, 21-24 APRIL 1997, pages 1275-1278 vol.2, XP002093539 ISBN 0-8186-7919-0, 1997, Los Alamitos, CA, USA, IEEE Comput. Soc. Press, USA * the whole document * --- | 1-15 | G10L5/06 |
| A | LI LEE ET AL: "Speaker normalization using efficient frequency warping procedures" 1996 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING CONFERENCE PROCEEDINGS (CAT. NO.96CH35903), 1996 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING CONFERENCE PROCEEDINGS, ATLANTA, GA, USA, 7-10 M, pages 353-356 vol. 1, XP002093540 ISBN 0-7803-3192-3, 1996, New York, NY, USA, IEEE, USA * the whole document * --- -/-- | 1-4, 10-12 | TECHNICAL FIELDS SEARCHED (Int.Cl.6) G10L |

The present search report has been drawn up for all claims

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| THE HAGUE | 16 February 1999 | Wanzeele, R |

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

## ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 98 10 4599

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

16-02-1999

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| US 5579436 A | 26-11-1996 | CA 2089903 A,C | 03-09-1993 |
| | | EP 0559349 A | 08-09-1993 |
| | | JP 6012093 A | 21-01-1994 |
| JP 08211887 A | 20-08-1996 | US 5742928 A | 21-04-1998 |

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82